

Characterising Probability Distributions via Entropies

Satyajit Thakor[†], Terence Chan[‡] and Alex Grant^{*}
 Indian Institute of Technology Mandi[†]
 University of South Australia[‡]
 Myriota Pty Ltd^{*}

Abstract—Characterising the capacity region for a network can be extremely difficult, especially when the sources are dependent. Most existing computable outer bounds are relaxations of the Linear Programming bound. One main challenge to extend linear program bounds to the case of correlated sources is the difficulty (or impossibility) of characterising arbitrary dependencies via entropy functions. This paper tackles the problem by addressing how to use entropy functions to characterise correlation among sources.

I. INTRODUCTION

This paper begins with a very simple and well known result. Consider a binary random variable X such that

$$p_X(0) = p \text{ and } p_X(1) = 1 - p.$$

While the entropy of X does not determine exactly what the probabilities of X are, it essentially determines the probability distribution (up to relabelling). To be precise, let $0 \leq q \leq 1/2$ such that $H(X) = h_b(q)$ where

$$h_b(q) \triangleq -q \log q - (1 - q) \log(1 - q).$$

Then either $p = q$ or $p = 1 - q$. Furthermore, the two possible distributions can be obtained from each other by renaming the random variable outcomes appropriately. In other words, there is a one-to-one correspondence between entropies and distribution (when the random variable is binary).

The basic question now is: *How “accurate” can entropies specify the distribution of random variables?* When X is not binary, the entropy $H(X)$ alone is not sufficient to characterise the probability distribution of X . In [1], it was proved that if X is a random scalar variable, its distribution can still be determined by using auxiliary random variables *subject to alphabet cardinality constraint*. The results can also be extended to random vector if the distribution is positive. However, the proposed approach cannot be generalised to the case when the distribution is not positive. In this paper, we take a different approach and generalise the result to any random vectors. Before we continue answering the question, we will briefly describe an application (based on network coding problems) of characterising distributions (and correlations) among random variables by using entropies.

Let the directed acyclic graph $\mathcal{G} = (\mathcal{V}, \mathcal{E})$ serve as a simplified model of a communication network with error-free point-to-point communication links. Edges $e \in \mathcal{E}$ have finite capacity $C_e > 0$. Let \mathcal{S} be an index set for a number of

multicast sessions, and $\{Y_s : s \in \mathcal{S}\}$ be the set of source random variables. These sources are available at the nodes identified by the mapping (a source may be available at multiple nodes) $a : \mathcal{S} \mapsto 2^{\mathcal{V}}$. Similarly, each source may be demanded by multiple sink nodes, identified by the mapping $b : \mathcal{S} \mapsto 2^{\mathcal{V}}$. For all s assume that $a(s) \cap b(s) = \emptyset$. Each edge $e \in \mathcal{E}$ carries a random variable U_e which is a function of incident edge random variables and source random variables.

Sources are i.i.d. sequences $\{(Y_s^n, s \in \mathcal{S}), n = 1, 2, \dots\}$. Hence, each $(Y_s^n, s \in \mathcal{S})$ has the same joint distribution, and is independent across different n . For notation simplicity, we will use $(Y_s, s \in \mathcal{S})$ to denote a generic copy of the sources at any particular time instance. However, within the same “time” instance n , the random variables $(Y_s^n, s \in \mathcal{S})$ may be correlated. We assume that the distribution of $(Y_s, s \in \mathcal{S})$ is known.

Roughly speaking, a link capacity tuple $\mathbf{C} = (C_e : e \in \mathcal{E})$ is achievable if one can design a network coding solution to transmit the sources $\{(Y_s^n, s \in \mathcal{S}), n = 1, 2, \dots\}$ to their respective destinations such that 1) the probability of decoding error is vanishing (as n goes to infinity), and 2) the number of bits transmitted on the link $e \in \mathcal{E}$ is at most nC_e . The set of all achievable link capacity tuples is denoted by \mathcal{R} .

Theorem 1 (Outer bound [2]): For a given network, consider the set of correlated sources $(Y_s, s \in \mathcal{S})$ with underlying probability distribution $P_{Y_{\mathcal{S}}}(\cdot)$. Construct any auxiliary random variables $(K_i, i \in \mathcal{L})$ by choosing a conditional probability distribution function $P_{K_{\mathcal{L}}|Y_{\mathcal{S}}}(\cdot)$. Let \mathcal{R}' be the set of all link capacity tuples $\mathbf{C} = (C_e : e \in \mathcal{E})$ such that there exists a polymatroid h satisfying the following constraints

$$h(X_{\mathcal{W}}, J_{\mathcal{Z}}) - H(Y_{\mathcal{W}}, K_{\mathcal{Z}}) = 0 \quad (1)$$

$$h(U_e | X_s : a(s) \rightarrow e, U_f : f \rightarrow e) = 0 \quad (2)$$

$$h(Y_s : u \in b(s) | X_{s'} : u \in a(s'), U_e : e \rightarrow u) = 0 \quad (3)$$

$$C_e - h(U_e) \geq 0 \quad (4)$$

for all $\mathcal{W} \subseteq \mathcal{S}, \mathcal{Z} \subseteq \mathcal{L}, e \in \mathcal{E}, u \in b(s)$ and $s \in \mathcal{S}$. Then

$$\mathcal{R} \subseteq \mathcal{R}' \quad (5)$$

where the notation $x \rightarrow y$ means x is incident to y and x, y can be an edge or a node.

Remark 1: The region \mathcal{R}' will depend on how we choose the auxiliary random variables $(K_i, i \in \mathcal{L})$. In the following, we give an example to illustrate this fact.

Consider the following network coding problem depicted

in Figure 1, in which three correlated sources Y_1, Y_2, Y_3 are available at node 1 and are demanded at nodes 3, 4, 5 respectively. Here, Y_1, Y_2, Y_3 are defined such that $Y_1 = (b_0, b_1)$, $Y_2 = (b_0, b_2)$ and $Y_3 = (b_1, b_2)$ for some independent and uniformly distributed binary random variables b_0, b_1, b_2 . Furthermore, the edges from node 2 to nodes 3, 4, 5 have sufficient capacity to carry the random variable U_1 available at node 2.

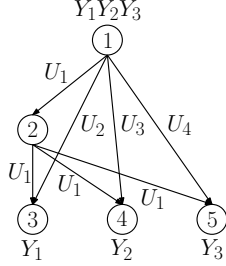


Fig. 1. A network example [2].

We consider two outer bounds obtained from Theorem 1 for the above network coding problem. In the first scenario, we use no auxiliary random variables, while in the second scenario, we use three auxiliary random variables such that

$$K_0 = b_0, K_1 = b_1, K_2 = b_2.$$

Let \mathcal{R}_i be respectively the outer bounds for the two scenarios. Then \mathcal{R}_2 is a proper subset of \mathcal{R}_1 . In particular, the link capacity tuple $(C_e = 1, e = 1, \dots, 4)$ is in the region $\mathcal{R}_1 \setminus \mathcal{R}_2$ [2]. This example shows that by properly choosing auxiliary random variables, one can better capture the correlations among the sources, leading to a strictly tighter/better outer bound for network coding. Construction of auxiliary random variables from source correlation was also considered in [3] to improve cut-set bounds.

II. MAIN RESULTS

In this section, we will show that by using auxiliary random variables, the probability distribution of a set of random variables (or a random vector) can be uniquely characterised from the entropies of these variables.

A. Random Scalar Case

Consider any ternary random variable X . Clearly, entropies of X and probability distributions are not in one-to-one correspondence. In [1], auxiliary random variables are used to in order to exactly characterise the distribution.

Suppose X is ternary, taking values from the set $\{1, 2, 3\}$. Suppose also that $p_X(x) > 0$ for all $x \in \{1, 2, 3\}$. Define random variables A_1, A_2 and A_3 such that

$$A_i = \begin{cases} 1 & \text{if } X = i \\ 0 & \text{otherwise.} \end{cases} \quad (6)$$

Clearly,

$$H(A_i|X) = 0, \quad (7)$$

$$H(A_i) = h_b(p_X(i)). \quad (8)$$

Let us further assume that $p_X(i) \leq 1/2$ for all i . Then by (8) and strict monotonicity of $h_b(q)$ in the interval $[0, 1/2]$, it seems at the first glance that the distribution of X is uniquely specified by the entropies of the auxiliary random variables.

However, there is a catch in the argument – The auxiliary random variables chosen are not arbitrary. When we “compute” the probabilities of X from the entropies of the auxiliary random variables, it is assumed to know how the random variables are constructed. Without knowing the “construction”, it is unclear how to find the probabilities of X from entropies.

More precisely, suppose we only know that there exists auxiliary random variables A_1, A_2, A_3 such that (7) and (8) hold (without knowing that the random variables are specified by (6)). Then we cannot determine precisely what the distribution of X is. Despite this complexity, [1], [2] showed a construction of auxiliary random variables from which the probability distribution can be characterised from entropies. The results will also be briefly restated as a necessary prerequisite for the vector case.

Let X be a random variable with support $\mathcal{N}_n = \{1, \dots, n\}$ and Ω be the set of all nonempty binary partitions of \mathcal{N}_n . In other words, Ω is the collection of all sets $\{\alpha, \alpha^c\}$ such that $\alpha \subseteq \mathcal{N}_n$, and both $|\alpha|$ and $|\alpha^c|$ are nonzero. We will use $\langle \alpha \rangle$ to denote the set $\{\alpha, \alpha^c\}$. To simplify notations, we may assume without loss of generality that α is a subset of $\{2, \dots, n\}$. Clearly, $|\Omega| = 2^{n-1} - 1$. Unless explicitly stated otherwise, we may assume without loss of generality that the probability that $X = i$ (denoted by p_i) is monotonic decreasing. In other words,

$$p_1 \geq \dots \geq p_n > 0.$$

Definition 1 (Partition Random Variables): A random variable X with support \mathcal{N}_n induces $2^{n-1} - 1$ random variables $A_{\langle \alpha \rangle}$ for $\alpha \in \Omega$ such that

$$A_{\langle \alpha \rangle} \triangleq \begin{cases} \alpha & \text{if } X \in \alpha \\ \alpha^c & \text{otherwise.} \end{cases} \quad (9)$$

We called $\{A_{\langle \alpha \rangle}, \alpha \in \Omega\}$ the collection of *binary partition random variables* of X .

Remark 2: If $|\alpha| = 1$ or $n - 1$, then there exists an element $i \in \mathcal{X}$ such that $A_{\langle \alpha \rangle} = \{i\}$ if and only if $X = i$. Hence, $A_{\langle \alpha \rangle}$ is essentially a binary variable *indicating/detecting whether* $X = i$ *or not*. As such, we call $A_{\langle \alpha \rangle}$ an *indicator variable*. Furthermore, when $n \geq 3$, there are exactly n indicator variables, one for each element in \mathcal{N}_n .

Theorem 2 (Random Scalar Case): Suppose X is a random variable with support \mathcal{N}_n . For any $\langle \alpha \rangle \in \Omega$, let $A_{\langle \alpha \rangle}$ be the corresponding binary partition random variables. Now, suppose X^* is another random variable such that 1) the size of its support \mathcal{X}^* is at most the same as that of X , and 2) there exists random variables $(B_{\langle \alpha \rangle}, \alpha \in \Omega)$ satisfying the following conditions:

$$H(B_{\langle \alpha \rangle}, \alpha \in \Delta) = H(A_{\langle \alpha \rangle}, \alpha \in \Delta) \quad (10)$$

$$H(B_{\langle \alpha \rangle}|X^*) = 0 \quad (11)$$

for all $\Delta \subseteq \Omega$. Then there is a mapping

$$\sigma : \mathcal{N}_n \rightarrow \mathcal{X}^*$$

TABLE I
PROBABILITY DISTRIBUTIONS OF X AND X^*

		X_2			
		1	2	3	4
X_1	a	1/8	1/8	0	0
	b	1/8	1/8	0	0
	c	0	0	1/8	1/8
	d	0	0	1/8	1/8

		X_2^*			
		1	2	3	4
X_1^*	a	1/8	1/8	0	0
	b	0	1/8	1/8	0
	c	0	0	1/8	1/8
	d	1/8	0	0	1/8

such that $\Pr(X = i) = \Pr(X^* = \sigma(i))$. In other words, the probability distributions of X and X^* are essentially the same (via renaming outcomes).

Proof: A sketch of the proof is shown in Appendix A. ■

B. Random Vector Case

Extension of Theorem 2 to the case of random vector has also been considered briefly in our previous work [1]. However, the extension is fairly limited in that work – the random vector must have a positive probability distribution and each individual random variable must take at least three possible values. In this paper, we overcome these restrictions and fully generalise Theorem 2 to the random vector case.

Example 1: Consider two random vectors $X = (X_1, X_2)$ and $X^* = (X_1^*, X_2^*)$ with probability distributions given in Table I. If we compare the joint probability distributions of X and X^* , they are different from each other. Yet, if we treat X and X^* as scalars (by properly renaming), then they indeed have the same distribution (both uniformly distributed over a support of size 8). This example shows that we cannot directly apply Theorem 2 to the random vector case, by simply mapping a vector into a scalar.

Theorem 3 (Random Vector): Suppose $X = (X_1, \dots, X_n)$ is a random vector with support \mathcal{X} of size at least 3. Again, let Ω be the set of all nonempty binary partitions of \mathcal{X} and $A_{\langle\alpha\rangle}$ be the binary partition random variable of X such that

$$A_{\langle\alpha\rangle} = \begin{cases} \alpha & \text{if } X \in \alpha \\ \alpha^c & \text{otherwise} \end{cases} \quad (12)$$

for all $\langle\alpha\rangle \in \Omega$.

Now, suppose $X^* = (X_1^*, \dots, X_M^*)$ is another random vector where there exists random variables

$$(B_{\langle\alpha\rangle}, \langle\alpha\rangle \in \Omega)$$

such that for any subset Δ of Ω and $\tau \subseteq \{1, \dots, M\}$,

$$H(B_{\langle\alpha\rangle}, \langle\alpha\rangle \in \Delta, X_j^*, j \in \tau) = H(A_{\langle\alpha\rangle}, \langle\alpha\rangle \in \Delta, X_j, j \in \tau). \quad (13)$$

Then the joint probability distributions of $X = (X_1, \dots, X_n)$ and $X^* = (X_1^*, \dots, X_n^*)$ are essentially the same. More pre-

cisely, there exists bijective mappings σ_m for $m = 1, \dots, M$ such that

$$\Pr(X = (x_1, \dots, x_M)) = \Pr(X^* = (\sigma_1(x_1), \dots, \sigma_M(x_M))). \quad (14)$$

Proof: See Appendix B. ■

C. Applications: Network coding outer bounds

Together with Theorem 1 and the characterisation of random variable using entropies, we obtain the following outer bound \mathcal{R}'' on the set of achievable capacity tuples.

Corollary 1: For any given network, consider the set of correlated sources $(Y_s, s \in \mathcal{S})$ with underlying probability distribution $P_{Y_S}(\cdot)$. From this distribution, construct binary partition random variables $A_{\langle\alpha\rangle}^{\mathcal{W}}$ for every subset $\mathcal{W} \subseteq \mathcal{S}$ as described in Theorem 1 (for scalar subsets) and Theorem 3 (for vector subsets). Let \mathcal{R}'' be the set of all link capacity tuples $\mathbf{C} = (C_e : e \in \mathcal{E})$ such that there exists an almost entropic function $h \in \overline{\Gamma}^*$ satisfying the constraints (2)-(4) and

$$h(X_{\mathcal{W}}, B_{\langle\alpha\rangle}^{\mathcal{W}}) - H(Y_{\mathcal{W}}, A_{\langle\alpha\rangle}^{\mathcal{W}}) = 0 \quad (15)$$

for every $\mathcal{W} \subseteq \mathcal{S}, \langle\alpha\rangle \in \Omega, e \in \mathcal{E}, u \in b(s)$ and $s \in \mathcal{S}$. Then $\mathcal{R} \subseteq \mathcal{R}''$. Replacing $\overline{\Gamma}^*$ by Γ , we obtain an *explicitly computable outer bound* $\mathcal{R}''(\Gamma)$.

III. CONCLUSION

In this paper, we showed that by using auxiliary random variables, entropies are sufficient to uniquely characterise the probability distribution of a random vector (up to outcome relabelling). Yet, there are still many open questions remained to be answered. For example, the number of auxiliary random variables used are exponential to the size of the support. Can we reduce the number of auxiliary random variables? What is the tradeoff between the number of auxiliary variables used and the quality of how well entropies can characterise the distribution? To the extreme, if only one auxiliary random variable can be used, how can one pick the variable to best describe the distribution?

REFERENCES

- [1] S. Thakor, T. Chan, and A. Grant, "Characterising correlation via entropy functions," in *Information Theory Workshop (ITW), 2013 IEEE*, pp. 1–2, Sept 2013.
- [2] S. Thakor, T. Chan, and A. Grant, "Bounds for network information flow with correlated sources," in *Australian Communications Theory Workshop (AusCTW)*, (Melbourne, Australia), pp. 43–48, Feb. 2011.
- [3] A. Gohari, S. Yang, and S. Jaggi, "Beyond the cut-set bound: Uncertainty computations in network coding with correlated sources," *IEEE Trans. Inform. Theory*, vol. 59, pp. 5708–5722, Sept 2013.

APPENDIX A SCALAR CASE

The main ingredients in the proofs for Theorems 2 and 3 are the properties of the partition random variables, which will be reviewed as follows. By understanding the properties, we can better understand the logic behind Theorem 2.

Lemma 1 (Properties): Let X be a random variable with support \mathcal{N}_n , and $(A_{\langle\alpha\rangle}, \alpha \in \Omega)$ be its induced binary partition random variables. Then the following properties hold:

- 1) (Distinctness) For any $\langle\alpha\rangle \neq \langle\beta\rangle$,

$$H(A_{\langle\alpha\rangle}|A_{\langle\beta\rangle}) > 0, \quad (16)$$

$$H(A_{\langle\beta\rangle}|A_{\langle\alpha\rangle}) > 0. \quad (17)$$

- 2) (Completeness) Let A^* be a binary random variable such that $H(A^*|X) = 0$ and $H(A^*) > 0$. Then there exists $\langle\alpha\rangle \in \Omega$ such that

$$H(A^*|A_{\langle\alpha\rangle}) = H(A_{\langle\alpha\rangle}|A^*) = 0. \quad (18)$$

In other words, $A_{\langle\alpha\rangle}$ and A^* are essentially the same.

- 3) (Basis) Let $\langle\alpha\rangle \in \Omega$. Then there exists

$$\langle\beta_1\rangle, \dots, \langle\beta_{n-2}\rangle \in \Omega$$

such that

$$H(A_{\langle\beta_k\rangle}|A_{\langle\alpha\rangle}, A_{\langle\beta_1\rangle}, \dots, A_{\langle\beta_{k-1}\rangle}) > 0 \quad (19)$$

for all $k = 1, \dots, n-2$.

Among all binary partition random variables, we are particularly interested in those indicator random variables. The following proposition can be interpreted as “entropic characterisation” for those indicator random variables.

Proposition 1 (Characterising indicators): Let X be a random variable of support \mathcal{N}_n where $n \geq 3$. Consider the binary partition random variables induced by X . Then for all $i \geq 2$,

- 1) $H(A_{\langle i \rangle}|A_{\langle j \rangle}, j > i) > 0$, and
2) For all $\alpha \in \Omega$ such that $H(A_{\langle\alpha\rangle}|A_{\langle j \rangle}, j > i) > 0$,

$$H(A_{\langle i \rangle}) \leq H(A_{\langle\alpha\rangle}). \quad (20)$$

- 3) Equalities (20) hold if and only if $A_{\langle\alpha\rangle}$ is an indicator random variable detecting an element $\ell \in \mathcal{N}_n$ such that

$$p_\ell = p_i.$$

- 4) Let $\beta \subseteq \{2, \dots, n\}$. The indicator random variable $A_{\langle 1 \rangle}$ is the only binary partition variable of X such that

$$H(A_{\langle\alpha\rangle}|A_{\langle j \rangle}, j \in \beta) > 0$$

for all proper subset β of $\{2, \dots, n\}$.

Sketch of Proof for Theorem 2: Let X be a random scalar and $A_{\langle\alpha\rangle}$ for $\langle\alpha\rangle \in \Omega$ are its induced partition random variables. Suppose X^* is another random variable such that 1) the size of its support \mathcal{X}^* is at most the same as that of X , and 2) there exists random variables $(B_{\langle\alpha\rangle}, \alpha \in \Omega)$ satisfying (10) and (11).

Roughly speaking, (10) and (11) mean that the set of random variables $(B_{\langle\alpha\rangle}, \alpha \in \Omega)$ satisfy most properties as ordinary partition random variables. To prove the theorem, our first immediate goal is to prove that those random variables $B_{\langle\alpha\rangle}$ are indeed binary partition random variables. In particular, we can prove that

- 1) (Distinctness) All the random variables $B_{\langle\alpha\rangle}$ for $\langle\alpha\rangle \in \Omega$ are distinct and have non-zero entropies.
2) (Basis) Let $\langle\alpha\rangle \in \Omega$. Then there exists

$$\langle\beta_1\rangle, \dots, \langle\beta_{n-2}\rangle \in \Omega$$

such that

$$H(B_{\langle\beta_k\rangle}|B_{\langle\alpha\rangle}, B_{\langle\beta_1\rangle}, \dots, B_{\langle\beta_{k-1}\rangle}) > 0 \quad (21)$$

for all $k = 1, \dots, n-2$.

- 3) (Binary properties) For any $\langle\alpha\rangle \in \Omega$, $B_{\langle\alpha\rangle}$ is a binary partition random variable of X^* . In this case, we may assume without loss of generality that there exists $\omega_{\langle\alpha\rangle} \subseteq \mathcal{X}^*$ such that

$$B_{\langle\alpha\rangle} = \begin{cases} \omega_{\langle\alpha\rangle} & \text{if } X^* \in \omega_{\langle\alpha\rangle} \\ \omega_{\langle\alpha\rangle}^c & \text{otherwise.} \end{cases} \quad (22)$$

- 4) (Completeness) Let B^* be a binary partition random variable of X^* with non-zero entropy. Then there exists $\langle\alpha\rangle \in \Omega$ such that

$$H(B^*|B_{\langle\alpha\rangle}) = H(B_{\langle\alpha\rangle}|B^*) = 0. \quad (23)$$

Then by (10) – (11) and Proposition 1, we show that $B_{\langle\alpha\rangle}$ satisfies all properties which are only satisfied by the indicator random variables. Thus, we prove that $B_{\langle\alpha\rangle}$ is an *indicator variable* if $|\alpha| = 1$. Finally, once we have determined which are the indicator variables, we can immediately determine the probability distribution. As $H(A_{\langle\alpha\rangle}) = H(B_{\langle\alpha\rangle})$ for all $\langle\alpha\rangle \in \Omega$, the distribution of X^* is indeed the same as that of X (subject to relabelling). ■

APPENDIX B VECTOR CASE

In this appendix, we will sketch the proof for Theorem 3, which extends Theorem 2 to the random vector case.

Consider a random vector

$$X = (X_m : m \in \mathcal{N}_M). \quad (24)$$

We will only consider the general case¹ where the support size of X is at least 3, i.e., $\mathcal{S}(X_m : m \in \mathcal{N}_M) \geq 3$.

Let \mathcal{X} be the support of X . Hence, elements of \mathcal{X} is of the form $x = (x_1, \dots, x_M)$ such that

$$\Pr(X_m = x_m, m \in \mathcal{N}_M) > 0$$

if and only if $x \in \mathcal{X}$.

The collection of binary partition random variables induced by the random vector $X = (X_m, m \in \mathcal{N}_M)$ is again indexed by $(A_{\langle\alpha\rangle}, \langle\alpha\rangle \in \Omega)$. As before, we may assume without loss of generality that

$$A_{\langle\alpha\rangle} = \begin{cases} \alpha & \text{if } X \in \alpha \\ \alpha^c & \text{otherwise.} \end{cases} \quad (25)$$

Now, suppose $(B_{\langle\alpha\rangle}, \langle\alpha\rangle \in \Omega)$ is a set of random variables satisfying the properties as specified in Theorem 3. Invoking Theorem 2 (by treating the random vector X^* as one discrete variable), we can prove the following.

- 1) The size of the support of X^* and X are the same.
2) $B_{\langle\alpha\rangle}$ is a binary partition variable for all $\langle\alpha\rangle \in \Omega$.
3) The set of variables $(B_{\langle\alpha\rangle}, \langle\alpha\rangle \in \Omega)$ contains all distinct binary partition random variables induced by X^* .

¹ In the special case when the support size of X is less than 3, the theorem can be proved directly.

4) $B_{\langle x \rangle}$ is an indicator variable for all $x \in \mathcal{X}$.

According to definition, $A_{\langle x \rangle}$ is defined as an indicator variable for detecting x . However, while $B_{\langle x \rangle}$ is an indicator variable, the subscript x in $B_{\langle x \rangle}$ is only an index. The element detected by $B_{\langle x \rangle}$ can be any element in the support of X^* , which can be completely different from \mathcal{X} . To highlight the difference, we define the mapping σ such that for any $x \in \mathcal{X}$, $\sigma(x)$ is the element in the support of X^* that is detected by $B_{\langle x \rangle}$. In other words

$$A_{\langle \sigma(x) \rangle}^* = B_{\langle x \rangle}. \quad (26)$$

The following lemma follows from Theorem 2.

Lemma 2: For all $x \in \mathcal{X}$,

$$\Pr(X = x) = \Pr(X^* = \sigma(x)).$$

Let \mathcal{X}^* be the support of X^* . We similarly define Ω^* as the collection of all sets of the form $\{\gamma, \gamma^c\}$ where γ is a subset of \mathcal{X}^* and the sizes of γ and γ^c are non-zero. Again, we will use $\langle \gamma \rangle$ to denote the set $\{\gamma, \gamma^c\}$ and define

$$A_{\langle \gamma \rangle}^* = \begin{cases} \gamma & \text{if } X^* \in \gamma \\ \gamma^c & \text{otherwise.} \end{cases} \quad (27)$$

For any $\langle \alpha \rangle \in \Omega$, $B_{\langle \alpha \rangle}$ is a binary partition random variable of X^* . Hence, we may assume without loss of generality that there exists γ such that $A_{\langle \gamma \rangle}^* = B_{\langle \alpha \rangle}$. For notation simplicity, we may further extend² the mapping σ such that $A_{\langle \sigma(\alpha) \rangle}^* = B_{\langle \alpha \rangle}$ for all $\alpha \subseteq \mathcal{X}$.

Proposition 2: Let $\langle \alpha \rangle \in \Omega$. Suppose $A_{\langle \beta \rangle}$ satisfies the following properties:

- 1) For any $\gamma \subseteq \alpha$, $H(A_{\langle \beta \rangle} | A_{\langle \gamma \rangle}, x \in \gamma) = 0$ if and only if $\gamma = \alpha$.
- 2) For any $\gamma \subseteq \alpha^c$, $H(A_{\langle \beta \rangle} | A_{\langle \gamma \rangle}, x \in \gamma) = 0$ if and only if $\gamma = \alpha^c$.

Then $A_{\langle \beta \rangle} = A_{\langle \alpha \rangle}$.

Proof: Direct verification. ■

By definition of $B_{\langle \alpha \rangle}$ and Proposition 2, we have the following result.

Proposition 3: Let $\langle \alpha \rangle \in \Omega$. Then $B_{\langle \beta \rangle} = B_{\langle \alpha \rangle}$ is the only binary partition variable of X^* such that

- 1) For any $\gamma \subseteq \alpha$, $H(B_{\langle \beta \rangle} | B_{\langle \gamma \rangle}, x \in \gamma) = 0$ if and only if $\gamma = \alpha$.
- 2) For any $\gamma \subseteq \alpha^c$, $H(B_{\langle \beta \rangle} | B_{\langle \gamma \rangle}, x \in \gamma) = 0$ if and only if $\gamma = \alpha^c$.

Proposition 4: Let $\alpha \in \mathcal{X}$. Then $\langle \sigma(\alpha) \rangle = \langle \delta(\alpha) \rangle$, where $\delta(\alpha) = \{\sigma(x) : x \in \alpha\}$.

Proof: By Proposition 3, $B_{\langle \alpha \rangle} = A_{\langle \sigma(\alpha) \rangle}^*$ is the only variable such that

- 1) For any $\gamma \subseteq \alpha$, $H(A_{\langle \sigma(\alpha) \rangle}^* | A_{\langle \sigma(x) \rangle}^*, x \in \gamma) = 0$ if and only if $\gamma = \alpha$.
- 2) For any $\gamma \subseteq \alpha^c$, $H(A_{\langle \sigma(\alpha) \rangle}^* | A_{\langle \sigma(x) \rangle}^*, x \in \gamma) = 0$ if and only if $\gamma = \alpha^c$.

The above two properties can then be rephrased as

² Strictly speaking, $\sigma(\alpha)$ is not precisely defined. As $\langle \gamma \rangle = \langle \gamma^c \rangle$, $\sigma(\alpha)$ can either be γ or γ^c . Yet, the precise choice of $\sigma(\alpha)$ does not have any effects on the proof. However, we only require that when α is a singleton, $\sigma(\alpha)$ should also be a singleton.

1) For any $\delta(\gamma) \subseteq \delta(\alpha)$,

$$H(A_{\langle \sigma(\alpha) \rangle}^* | A_{\langle \sigma(x) \rangle}^*, \sigma(x) \in \delta(\gamma)) = 0$$

if and only if $\delta(\gamma) = \delta(\alpha)$

2) For any $\delta(\gamma) \subseteq \delta(\alpha^c)$,

$$H(A_{\langle \sigma(\alpha) \rangle}^* | A_{\langle \sigma(x) \rangle}^*, \sigma(x) \in \delta(\gamma)) = 0$$

if and only if $\delta(\gamma) = \delta(\alpha^c)$.

Now, we can invoke Proposition 2 and prove that $A_{\langle \delta(\alpha) \rangle}^* = A_{\langle \sigma(\alpha) \rangle}^*$ or equivalently, $\langle \delta(\alpha) \rangle = \langle \sigma(\alpha) \rangle$. The proposition then follows. ■

Proposition 5: Consider two distinct elements $x = (x_1, \dots, x_M)$ and $x' = (x'_1, \dots, x'_M)$ in \mathcal{X} . Let

$$\sigma(x) = y = (y_1, \dots, y_M) \quad (28)$$

$$\sigma(x') = y' = (y'_1, \dots, y'_M). \quad (29)$$

Then $x_m \neq x'_m$ if and only if $y_m \neq y'_m$.

Proof: First, we will prove the only-if statement. Suppose $x_m \neq x'_m$. Consider the following two sets

$$\Delta = \{x'' = (x''_1, \dots, x''_M) \in \mathcal{X} : x''_m \neq x_m\}, \quad (30)$$

$$\Delta^c = \{x'' = (x''_1, \dots, x''_M) \in \mathcal{X} : x''_m = x_m\}. \quad (31)$$

It is obvious that $H(A_{\langle \Delta \rangle} | X_m) = 0$. By (10)-(11), we have $H(B_{\langle \Delta \rangle} | X_m^*) = 0$. Hence, $B_{\langle \Delta \rangle} = A_{\langle \sigma(\Delta) \rangle}^*$. Since $H(B_{\langle \Delta \rangle} | X_m^*) = 0$, this implies $H(A_{\langle \sigma(\Delta) \rangle}^* | X_m^*) = 0$.

Now, notice that $x \in \Delta^c$ and $x' \in \Delta$. By Proposition 4, $\sigma(\Delta) = \{\sigma(x) : x \in \Delta\}$. Therefore, $y' = \sigma(x') \in \sigma(\Delta)$ and $y = \sigma(x) \notin \sigma(\Delta)$. Together with the fact that $H(A_{\langle \sigma(\Delta) \rangle}^* | X_m^*) = 0$, we can then prove that

$$y'_m \neq y''_m.$$

Next, we prove the if-statement. Suppose $y, y' \in \mathcal{X}^*$ such that $y_m \neq y'_m$. There exist x and x' such that (28) and (29) hold. Again, define

$$\Lambda \triangleq \{y'' = (y''_1, \dots, y''_M) \in \mathcal{X}^* : y''_m \neq y_m\}, \quad (32)$$

$$\Lambda^c \triangleq \{y'' = (y''_1, \dots, y''_M) \in \mathcal{X}^* : y''_m = y_m\}. \quad (33)$$

Then $H(A_{\langle \Lambda \rangle}^* | X_m^*) = 0$. Let $\Phi \triangleq \{x \in \mathcal{X} : \sigma(x) \in \Lambda\}$. By definition and Proposition 4, $B_{\langle \Phi \rangle} = A_{\langle \sigma(\Phi) \rangle}^* = A_{\langle \Lambda \rangle}^*$. Hence, we have $H(B_{\langle \Phi \rangle} | X_m^*) = 0$ and consequently $H(A_{\langle \Phi \rangle} | X_m) = 0$. On the other hand, it can be verified from definition that $x \in \Phi^c$ and $x' \in \Phi$. Together with that $H(A_{\langle \Phi \rangle} | X_m) = 0$, we prove that $x_m \neq x'_m$. The proposition then follows. ■

Proof of Theorem 3: A direct consequence of Proposition 5 is that there exists bijective mappings $\sigma_1, \dots, \sigma_M$ such that $\sigma(x) = (\sigma_1(x_1), \dots, \sigma_M(x_M))$. On the other hand, Theorem 2 proved that $\Pr(X = x) = \Pr(X^* = \sigma(x))$. Consequently,

$$\begin{aligned} \Pr(X_1 = x_1, \dots, X_M = x_M) \\ = \Pr(X_1^* = \sigma_1(x_1), \dots, X_M^* = \sigma_M(x_M)). \end{aligned} \quad (34)$$

Therefore, the joint distributions of $X = (X_1, \dots, X_M)$ and $X^* = (X_1^*, \dots, X_M^*)$ are essentially the same (by renaming x_m as $\sigma_m(x_m)$). ■